Hybrid Models for Automatic Speech Recognition: a Comparison of Classical ANN and Kernel Based Methods

Ana I. García-Moral, Rubén Solera-Ureña, Carmen Peláez-Moreno, and Fernando Díaz-de-María

Department of Signal Theory and Communications
EPS-Universidad Carlos III de Madrid
Avda. de la Universidad, 30, 28911-Leganés (Madrid), Spain
{aisabel,rsolera,carmen,fdiaz}@tsc.uc3m.es

Abstract. Support Vector Machines (SVMs) are state-of-the-art methods for machine learning but share with more classical Artificial Neural Networks (ANNs) the difficulty of their application to input patterns of non-fixed dimension. This is the case in Automatic Speech Recognition (ASR), in which the duration of the speech utterances is variable. In this paper we have recalled the hybrid (ANN/HMM) solutions provided in the past for ANNs and applied them to SVMs performing a comparison between them. We have experimentally assessed both hybrid systems with respect to the standard HMM-based ASR system, for several noisy environments. On the one hand, the ANN/HMM system provides better results than the HMM-based system. On the other, the results achieved by the SVM/HMM system are slightly lower than those of the HMM system. Nevertheless, such a results are encouraging due to the current limitations of the SVM/HMM system.

Key words: Robust ASR, Additive noise, Machine Learning, Hybrid systems, Artificial Neural Networks, Support Vector Machines, Hidden Markov Models

1 Introduction

Hidden Markov Models (HMMs) have become the most employed core technique for Automatic Speech Recognition (ASR). After several decades of intense research work in the field, it seems that HMM-based ASR systems are very close to reach their limit of performance. Some alternative approaches, most of them based on Artificial Neural Networks (ANNs), were proposed during the late eighties and early nineties. Among them, it is worth to mention hybrid ANN/HMM systems (see [1] for an overview), since the reported results were comparable or even slightly superior to those achieved by HMMs.

On the other hand, during the last decade, a new tool appeared in the field of machine learning that has proved to be able to cope with hard classification problems in several fields of application: the Support Vector Machines (SVMs)

[2]. The SVMs are effective discriminative classifiers with several outstanding characteristics, namely: their solution is that with maximum margin; they are capable to deal with samples of a very high dimensionality; and their convergence to the minimum of the associated cost function is guaranteed.

Nevertheless, it seems clear that the application of these kernel-based machines to the ASR problem is not straightforward. In our opinion, there are three main difficulties to overcome: 1) SVMs are originally static classifiers and have to be adapted to deal with the variability of duration of speech utterances; 2) the SVMs were originally formulated as binary classifiers while the ASR problem is multiclass; and 3) current SVM training algorithms are not able to manage the huge databases typically used in ASR. In order to cope with these difficulties, some researchers have suggested hybrid SVM/HMM systems [3, 4], that notably resemble the previous hybrid ANN/HMM systems ([5]). In this paper we comparatively describe both types of hybrid systems (SVM/ and ANN/HMM), highlighting both their common fundamentals and their special characteristics, and conduct an experimental performance comparison for both clean and noisy speech recognition tasks.

2 Hybrid systems for ASR

As a result of the difficulties found in the application of ANNs to speech recognition, mostly motivated by the duration variability of the speech instances corresponding to the same class, a variety of different architectures and novel training algorithms that combined both HMMs with ANNs were proposed in the late eighties and early nineties. For a comprehensive survey of these techniques see [1]. In this paper, we have focused on those that employ ANNs (and SVMs) to estimate the HMM state posterior probabilities proposed by Bourlard and Morgan ([5, 6]).

The starting point for this approach is the well-know property of using feed-forward networks such as multi-layer perceptrons (MLPs) for estimating *a-posteriori* probabilities given two conditions: 1) there must be high enough number of input samples to train a good approximation between the input and output layers; and 2) a global minimum error criterion must be used to train the network (for example, mean square error or relative entropy).

The fundamental advantage of this approach is that it introduces a discriminative technique (ANN) into a generative system (HMM) while retaining their ability to handle the temporal variability of the speech signal.

However, this original formulation had to be modified to estimate the true emission (likelihood) probabilities by applying the Bayes' rule. Therefore, the *a-posteriori* probabilities should be normalized by the class priors to obtain what is called *scaled likelihoods*. This fact was further reinforced by posterior theoretical developments in the search of a global ANN optimization procedure (see [7]).

Thus, systems of this type keep being locally discriminant given that the ANN was trained to estimate *a-posteriori* probabilities. However, it can also be shown that, in theory, HMMs can be trained using local posterior probabilities

as emission probabilities, resulting in models that are both locally and globally discriminant. The problem is that there are generally mismatches between the prior class probabilities implicit to the training data and the priors that are implicit to the lexical and syntactic models that are used in recognition. In fact, some experimental results show that for certain cases the division by the priors is not necessary [7].

Among the advantages of using hybrid approaches we highlight the following (from [7]):

- Model accuracy: both MLPs and SVMs have more flexibility to provide more accurate acoustic models including the possibility of using different combinations of features as well as different sizes of context.
- Local discrimination ability (at a frame level) provided by MLPs.
- Parsimonious use of parameters: all the classes share the same ANN parameters.
- HMMs and MLPs exhibit complementary abilities for ASR tasks, which lead to higher recognition rates.

3 Experimental Setup

3.1 Database

We have used the well-known SpeechDat Spanish database for the fixed telephone network [8]. This database comprises recordings from 4000 Spanish speakers recorded at 8 KHz over the fixed PSTN using an E-1 interface, in a noiseless office environment.

In our experiments we have used a large vocabulary (more than 24000 words) continuous speech recognition database. The training set contains approximately 50 hours of voice from 3146 speakers (71000 utterances). The callers spoke 40 items whose contents are varied, comprising isolated and connected digits, natural numbers, spellings, city and company names, common applications words, phonetically rich sentences, etc. Most items are read and some of them are spontaneously spoken. The test set, corresponding to a connected digits task, contains approximately 2122 utterances and 19855 digits (5 hours of voice) from 499 different speakers.

3.2 Parameterization

In our experiments we have used the classical parameterization based on 12 MFCCs (Mel-Frequency Cepstral Coefficients) plus energy, and the first and second derivatives. These MFCCs are computed every 10 ms using a time window of 25 ms. Thus, the resulting feature vectors have 39 components. In this work, we have considered a per-utterance normalization, that is, every parameter is normalized in mean and variance according to the following expression:

$$\hat{x}_i[n] = \frac{x_i[n] - \mu_f}{\sigma_f}, \tag{1}$$

where $x_i[n]$ represents the i^{th} component of the feature vector corresponding to frame n, μ_f is the estimated mean from the whole utterance, and σ_f is the estimated standard deviation. As a result, per-utterance normalization will be more appropriate in the case of noisy environments where training and testing conditions do not match.

3.3 Database contamination

We have tested our systems in clean conditions and in presence of additive noise. For that purpose, we have used two different kinds of noises (white and babble) extracted from the NOISEX database [9]. These noises have been added to the clean speech signals at four different signal-to-noise ratios (SNRs), namely: 12 dB, 9 dB, 6 dB and 3 dB. Only the testing subset has been corrupted in the way previously stated, whereas the acoustic models (GMMs (Gaussian Mixture Models) in the case of the baseline HMM system and the MLPs and the SVMs in the case of the hybrid systems) have been estimated or trained using only clean speech.

3.4 Baseline experiment with HMMs

The recognition rates achieved by a left-to-right HMM-based recognition system based in the COST-249 SpeechDat Reference Recognizer will be our reference results. We use 18 context-dependent phones (this is the number of phones usually used for digits recognition tasks in Spanish) with 3 states per phone. Emission probabilities for each state were modeled by a mixture of 32 Gaussians.

3.5 Experiments with Hybrid Recognition Systems

In this work we consider two different hybrid recognition systems, an ANN/HMM system and a SVM/HMM one. Both of them use a Viterbi decoder with *a-posteriori* probabilities as local scores as discussed in section 2.

The whole hybrid recognition system is composed of two stages. The first one estimates initial evidences for phones in the form of *a-posteriori* probabilities using an MLP or an SVM. The second stage consists of a classical Viterbi decoder where we replace the likelihoods estimates provided by the reference HMM-based recognition system by the posteriors obtained in the first stage.

For the hybrid systems presented in this paper, we have partitioned every phone into three segments. For this purpose, we have obtained a segmentation of the training database by performing a forced alignment with the HMM baseline system, considering each segment delimited by the state transitions. Experimental results with both ANN/HMM and SVM/HMM hybrid systems show significant improvements in the word recognition rate due to the use of three classes per phone, especially for the case of the SVM-based system (see [4]).

Whereas the reference HMM-based recognition system uses the whole training data set (71000 utterances), the hybrid SVM-based recognition system only

uses a small portion of the available training data, due to a practical limitation regarding the number of training samples that the SVM software can consider. Thus, in order to compare the two hybrid systems, we decided to use the same small quantity of training data in the ANN/HMM hybrid system, although we also obtained some results using the whole training data set. Therefore, we have considered useful to evaluate the evolution of the accuracy of each system performing incremental tests using balanced subsets of the available training data set (equal number of frames per class -three classes per phone-, randomly selected from the whole training set), between 250 and 20000 frames per class.

Experiments with ANNs *A posteriori* probabilities used by the Viterbi decoder are obtained using a MLP trained on either a smaller version of the training data set or the whole training data set, as we mentioned before. The MLP has one hidden layer with 1800 units. There are 39 input units corresponding to the feature vector dimension described in section 3.2, and 54 output units, each of them corresponding to one of the three parts of the 18 phones considered, as we described in section 3.5. The MLP is trained using the relative entropy criterion and the back-propagation factor μ was experimentally fixed at 0.02 by using a separate tuning set.

Experiments with SVMs In this case, a multiclass SVM (using the *1-vs-1* approach) is used to estimate posteriors for each frame using Platt's approximation ([10]). The SVM uses a RBF (Radial Basis Function) kernel whose parameter, γ, must be tuned by means of a cross-validation process, as well as a parameter C, which establishes a compromise between error minimization and generalization capability in the SVM. The values we have used in our experiments are $C = 2$ and $\gamma = 0.03125$ also obtained empirically using the tuning set we already mentioned in 3.5. For more details about the hybrid SVM/HMM system, please refer to [4].

4 Results and Discussion

This section is devoted to the presentation and discussion of the results obtained by the systems described in the previous section.

Preliminary experiments show a similar behaviour of both SVMs and ANNs at a frame classification level. We can also see in figure 1 that better results are achieved when more samples are added to the training database, up to a final frame recognition rate around 72% obtained for the maximum number of input samples that our SVM-based system can handle (1080000, 20000 frames per class). Nevertheless, when we use our MLP-based system, that can handle the whole (and not balanced) set of input samples (16378624 frames), we manage to improve this frame recognition rate up to 78.47%. Similar behaviour is expected for the SVM/HMM hybrid system if the employed software could process such an amount of input samples.

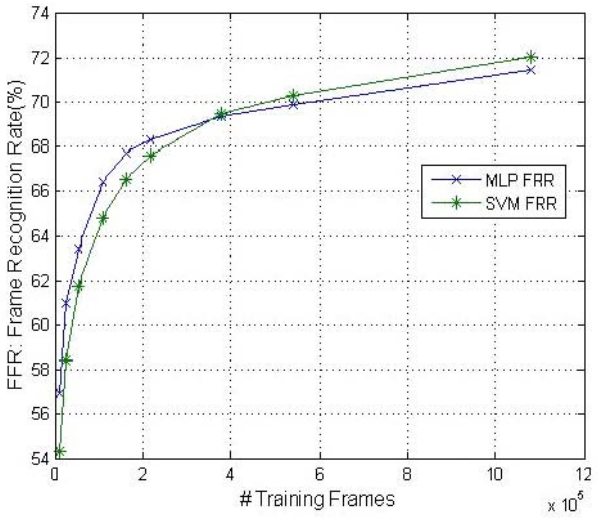

Fig. 1. *Frame recognition rate of ANNs and SVMs.*

We compare our hybrid systems to the standard HMM-based speech recognition system at word and sentence levels, in the different noise environments described in section 3.3. We can see in figures 2 and 3 the Word Recognition Rate (WRR) and the Sentence Recognition Rate (SRR), respectively, of the three systems. Thus, we can notice that, using only a 6.6% of the available data samples, our hybrid systems get results which are comparable or even better (in the case of the ANN/HMM system) than the standard HMM-based system trained using the entire database.

The SVM software used in the experiments [11], which requires to keep the kernel matrix in memory, is the responsible for the limit in the SVM/HMM training data set. However, without this restriction, we have observed in the ANN/HMM system that the more samples we add to the training database, the higher is the improvement in WRR and SRR with regard to the baseline HMM-based system, as we could see at the frame classification level.

In addition, as we have stated in Section 2, both SVMs and ANNs provide a-posteriori probabilities to the Viterbi decoder, whereas what we really need (and HMMs compute) are likelihoods [5]. We tried to use likelihoods in the ANN/HMM system, but we achieved worse results than that of posterior probabilities in all the cases except for the case we train the ANN/HMM hybrid using the entire database. If we analyze these results, we can see that, in fact, when we train the MLP using a balanced set of samples, the a-priori probability is the same for all the classes and posteriors provided by the MLP are actually scaled likelihoods.

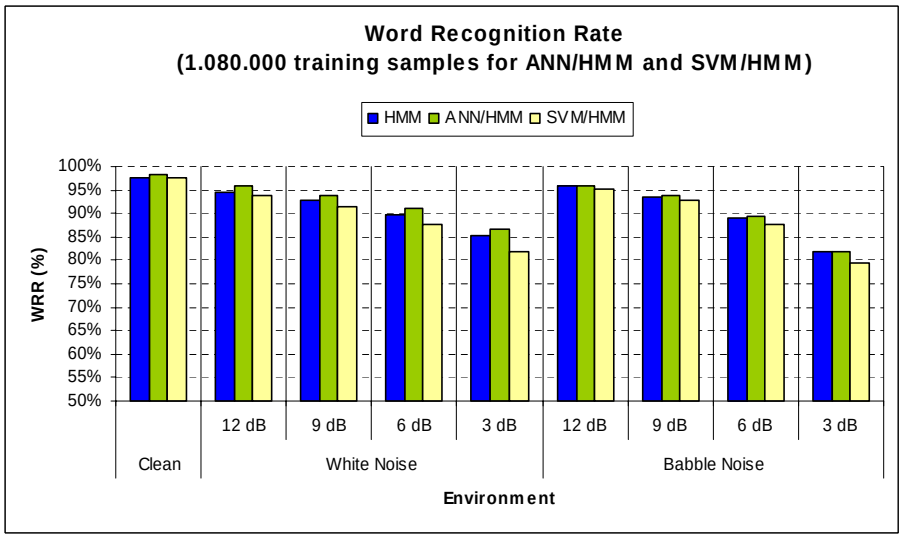

Fig. 2. *Word recognition rate of HMMs and Hybrid ANN- and SVM-based systems.*

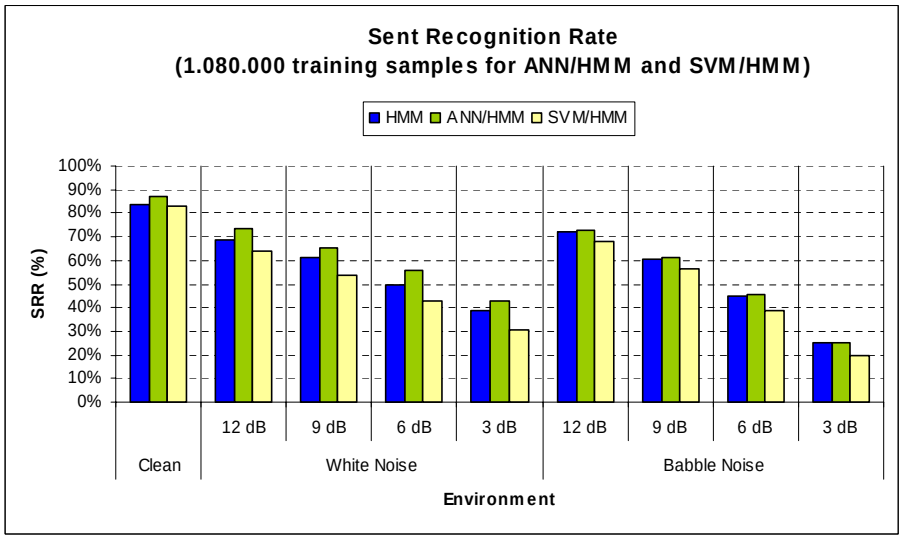

Fig. 3. *Sentence recognition rate of HMMs and Hybrid ANN- and SVM-based systems.*

5 Conclusions

The clear success of SVMs classifiers in several fields of application has called the attention of researchers in the field of ASR. The first attempts to use SVMs for connected-digit recognition have resulted in hybrid SVM/HMM systems [4] that resemble the hybrid systems based on ANN proposed during the last decade. Consequently, it becomes relevant to compare the performance achieved by both types of systems. Furthermore, since the robustness of the ASR systems is one of the current open problems, the comparative assessment of hybrid systems should be carried out in a noisy environment.

The ANN/HMM and SVM/HMM hybrid systems presented in this work are inspired in the work due to Bourlard and Morgan [5]. Our more significant contribution with respect to this reference consists of using sub-phone units. Specifically, three classes (parts) per phone are considered instead of one.

Some limitations regarding the publicly available software implementation of the SVMs have prevented us to train our hybrid SVM/HMM ASR system using the whole training set. Therefore, in order to carry out a fair comparison, the hybrid ANN/HMM has been trained using the same small subset of the training set. In this conditions, the achieved results can be summarized as follows:

- The hybrid ANN/HMM system provides slightly better results than the HMM-based system used as reference, for all the noise types and SNR values considered.
- The performance of the hybrid SVM/HMM system is slightly lower than that of the HMM-based system.

In our opinion these results are encouraging. On the one hand, the hybrid ANN/HMM system using sub-phone units turns out to be competitive. On the other hand, though the design of the hybrid SVM/HMM system is still preliminar, it has reached a reasonable level of performance. In particular, as can be expected from previous results in an isolated-digit recognition task [12], the maximum margin principle used for its training can make an important difference in noisy environment. For that purpose, several issues should be addressed; for example: the possibility to incorporate more training samples, the addition of a wider temporal context in the feature vectors and the selection of appropriate feature sets. Besides, hybrid systems are more amenable for its use with different types of parameterizations that do not comply with the restrictions of independence imposed by HMMs. This could result advantageous in the search of robustness.

Acknowledgments. This work is partially supported by the regional grant (Comunidad Autónoma de Madrid - UC3M) CCG06-UC3M/TIC-0812

References

1. Trentin, E., Gori, M.: A Survey of Hybrid ANN/HMM Models for Automatic Speech Recognition. Neurocomputing **37** (2001) 91–126

2. Boser, B.E., Guyon, I., Vapnik, V.: A Training Algorithm for Optimal Margin Classifiers. In: Computational Learning Theory. (1992) 144–152
3. Ganapathiraju, A., Hamaker, J., Picone, J.: Hybrid SVM/HMM Architectures for Speech Recognition. In: Proceedings of the 2000 Speech Transcription Workshop. Volume 4., Maryland (USA) (2000) 504–507
4. Padrell-Sendra, J., Martín-Iglesias, D., Díaz-de-María, F.: Support Vector Machines for Continuous Speech Recognition. In: Proceedings of the 14th European Signal Processing Conference, Florence (Italy) (2006)
5. Bourlard, H., Morgan, N.: Connectionist Speech Recognition: a Hybrid Approach. Boston: Kluwer Academic, Norwell, MA (USA) (1994)
6. Morgan, N., Bourlard, H.: Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach. IEEE Signal Processing Magazine (1995) 25–42
7. Bourlard, H., Morgan, N., Giles, C.L., Gori, M.: Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions. In: Adaptive Processing of Sequences and Data Structures. International Summer School on Neural Networks 'E. R. Caianiello'. Tutorial Lectures. Springer-Verlag, Germany; Berlin (1998) 389–417
8. Moreno, A.: SpeechDat Spanish Database for Fixed Telephone Network. Technical report, Technical University of Catalonia (1997)
9. Varga, A.P., Steenneken, J.M., Tolimson, M., Jones, D.: The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. Technical report, DRA Speech Research Unit (1992)
10. Platt, J.C.: Probabilities for SV Machines. In: Advances in Large Margin Classifiers. MIT Press (1999) 61–74
11. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at http://www.csie.ntu.edu.tw/cjlin/∼libsvm.
12. Solera-Ureña, R., Martín-Iglesias, D., Gallardo-Antolín, A., Peláez-Moreno, C., Díaz-de-María, F.: Robust ASR Using Support Vector Machines. Speech Communication **49(4)** (2007) 253–267